

团 体 标 准

T/SZAS 112—2026

大模型金融应用技术规范

Technical specification for financial applications of large models

2026-07-21 发布

2026-08-01 实施

深圳市标准化协会 发布

目 次

前言 II

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 2

5 技术框架 2

6 资源池要求 4

7 金融优化训练数据要求 4

8 金融应用大模型要求 6

9 系统组件要求 7

10 管控服务平台要求 9

11 大模型金融应用要求 10

参考文献 14

前 言

本文件按照 GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由深圳国家金融科技测评中心有限公司提出。

本文件由深圳市标准化协会归口。

本文件起草单位：深圳国家金融科技测评中心有限公司、北京银联金卡科技有限公司、深圳市福田区金融服务和风险防控中心、中国平安人寿保险股份有限公司、深圳壹账通智能科技有限公司、山证科技（深圳）有限公司、粤港澳大湾区数字经济研究院、中电信翼金科技有限公司、星环信息科技（上海）股份有限公司、招商证券股份有限公司、交通银行股份有限公司深圳分行、深圳市网安计算机安全检测技术有限公司、北京大成（杭州）律师事务所、深圳市金融科技协会、腾讯云计算（北京）有限责任公司、麒麟软件有限公司、杭州高新区（滨江）区块链与数据安全研究院、平安科技（深圳）有限公司、中山证券有限责任公司、诺安基金管理有限公司、深圳前海微众银行股份有限公司、深圳市链壳科技有限公司、上海市数字证书认证中心有限公司、深圳市标准化协会、北京新诚远航科技有限责任公司、鹏城实验室、上海大学、银雁科技服务集团股份有限公司、深圳海云安网络安全技术有限公司、深圳鼎智通讯股份有限公司、深圳市脉山龙信息技术股份有限公司。

本文件主要起草人：罗丰、许晋元、董晶、杨波、叶晓聪、袁荣婷、刘强军、黄榴勇、彭思思、蒋利兵、万敏、乔弘、翁俊峰、张乐、张炜聪、黄明辉、蒋玉长、贺恩锋、张扬、杨弘飞、史文彬、邹倩霞、丁艳霞、穆婉青、郭健、林舟驰、杨琛、陈华龙、张晓明、卢国彬、李海同、胡斐然、郑伯伟、谢云、龙军、郑灿铭、孙鹏程、陈思、罗振伟、檀革勤、王建、陈明、战茅、韩福海、魏遵博、夏文君、王健宗、张旭龙、颜世瑞、蔡泽娟、关为、武轩、李素游、高玉翔、李成博、马强、宗春雨、杨晶、但丹、满小婕、朱纯超、朱纯娜、相洋、孙东宁、张远健、李家菁、谢朝海、李志勇、雷德诚、欧阳伟权、余国凡、汪书福、陈晓宁。

大模型金融应用技术规范

1 范围

本文件规定了大模型金融应用的技术框架、资源池要求、金融优化训练数据要求、金融应用大模型要求、系统组件要求、管控服务平台要求和大模型金融应用要求。

本文件适用于开展大模型金融应用活动、利用大模型处理金融业务的金融机构，为大模型金融应用的设计、开发、部署、运行、管理和评估提供技术指导。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB 45438 网络安全技术 人工智能生成合成内容标识方法
GB/T 35273 信息安全技术 个人信息安全规范
GB/T 45225 人工智能 深度学习算法评估
GB/T 45288.1 人工智能 大模型 第1部分：通用要求
GB/T 45654 网络安全技术 生成式人工智能服务安全基本要求
GB/T 45674 网络安全技术 生成式人工智能数据标注安全规范
JR/T 0171 个人金融信息保护技术规范
JR/T 0221 人工智能算法金融应用评价规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

大模型训练推理资源 large model training and inference resources

大模型训练推理所必需的基础设施资源，主要包括算力资源、存储资源及网络资源。

3.2

金融优化训练数据 financial fine-tuning data

用于优化训练金融大模型，提升金融大模型特定金融应用处理能力的的数据。

3.3

生成合成数据 generate or synthesize data

根据数据内容、格式、内在逻辑等要求，通过一定规则生成或合成的数据，作为金融优化训练数据来源之一，用于金融大模型训练。

3.4

准入控制 access control

在模型、数据、应用等对象部署使用前，通过安全审查、合规评估等措施进行控制的活动，以阻止不符合要求的对象被使用。

3.5

金融应用大模型 large model for financial applications

服务于大模型金融应用所需的大模型集合，包括基础大模型、金融大模型和辅助大模型。

3.6

基础大模型 foundation large model

通过大规模数据预训练构建的通用模型。

3.7

金融大模型 financial large model

通过金融优化训练数据构建的金融垂域大模型。

3.8

智能体 agent

感知环境，并主动采取行动以实现特定目标的实体。

[来源：GB/Z 185.1-2026, 3.1]

3.9

关键词库 keyword database

包含不良违法信息等关键词，用于提高大模型应用输入层和输出层的安全性。

3.10

安全风险验证库 safety risk validation database

包含已知的注入攻击等漏洞，用于进行攻击识别或开展大模型金融应用安全风险评估。

4 缩略语

下列缩略语适用于本文件。

RDMA: 远程直接内存访问 (Remote Direct Memory Access)

API: 应用程序接口 (Application Programming Interface)

MCP: 模型上下文协议 (Model Context Protocol)

TTFT: 首个令牌生成时延 (Time to First Token)

TPOT: 单令牌生成时延 (Time Per Output Token)

E2E Latency: 端到端时延 (End-to-End Latency)

QPS: 每秒查询数 (Queries Per Second)

TPS: 每秒事务数 (Transactions Per Second)

5 技术框架

5.1 大模型金融应用技术框架由资源池、金融优化训练数据、金融应用大模型、系统组件、管控服务平台和大模型金融应用组成，如图 1 所示，各模块之间形成资源供给、训练样本供给、推理支撑、应用支撑、统一管控、内容输入及内容输出的关系。

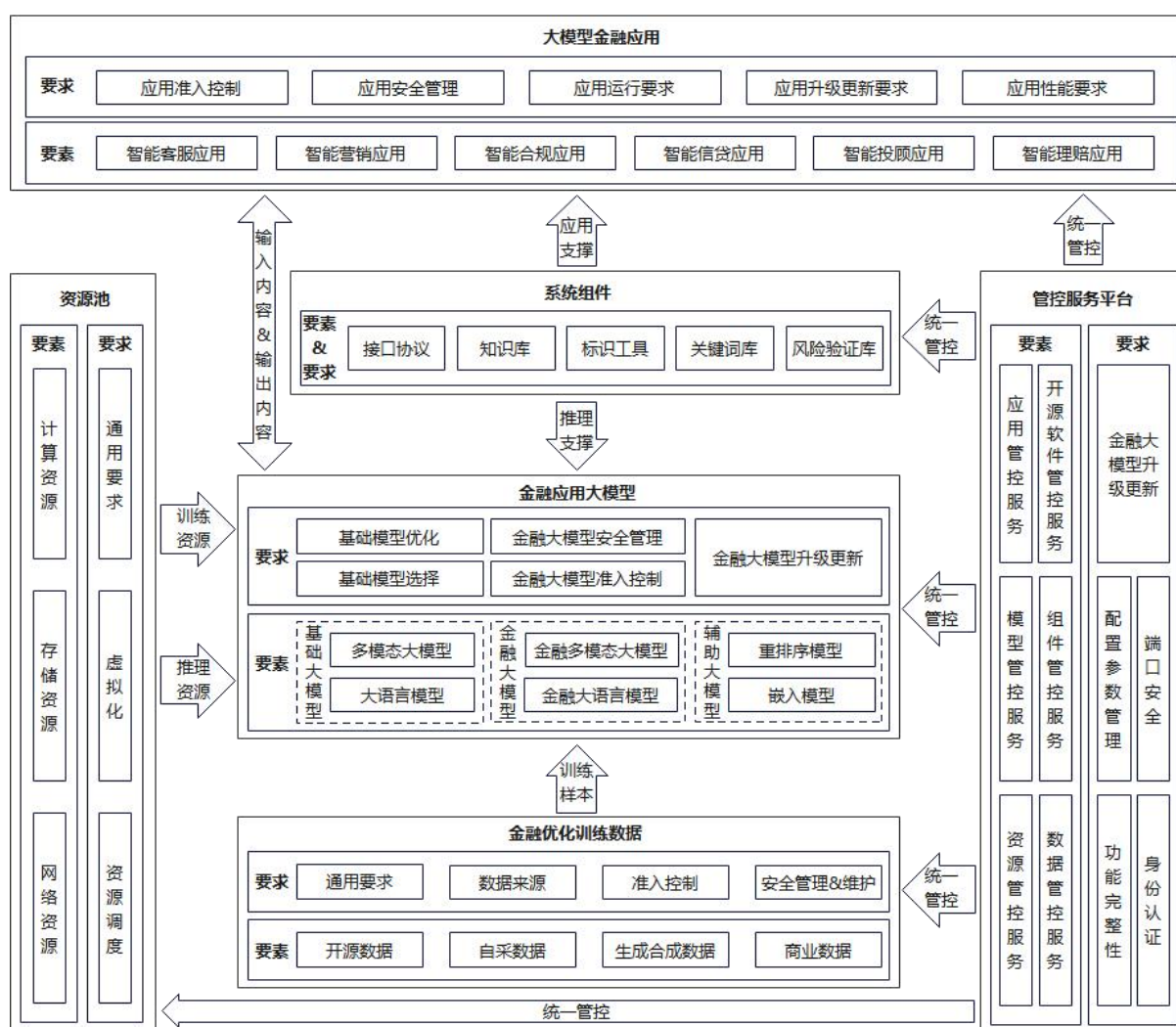


图 1 大模型金融应用技术框架

5.2 资源池是大模型金融应用所需的资源集合，包括计算资源、存储资源和网络资源，主要为金融应用大模型提供训练资源和推理资源。

5.3 金融优化训练数据是训练金融大模型所需的数据集合，包括开源数据、自采数据、生成合成数据和商业数据，主要为垂域金融大模型的训练提供所需样本。

5.4 金融应用大模型是大模型金融应用所需的大模型集合，包括基础大模型、金融大模型和辅助大模型，主要基于理解、生成、推理、检索、排序等能力为大模型金融应用提供高质量输出内容。

5.5 系统组件是大模型金融应用所需的组件集合，系统组件包括但不限于接口协议、知识库、标识工具、关键词库、风险验证库、沙箱、记忆系统、编排工具、推理引擎等，主要为大模型金融应用提供金融应用大模型以外的应用支撑能力，例如模型调用、知识增强、风险校验、输出控制等，同时也为金融应用大模型提供推理支撑。

5.6 管控服务平台是大模型金融应用的管控服务集合，包括资源管控服务、数据管控服务、模型管控服务、组件管控服务、开源软件管控服务和应用管控服务，主要提供满足本文件要求的准入控制、安全管理、升级更新、身份认证等统一管控服务能力。

5.7 大模型金融应用是服务于金融业务场景的应用集合，包括但不限于智能客服应用、智能营销应用、智能合规应用、智能信贷应用、智能投顾应用、智能理赔应用等，主要依托金融应用大模型为金融机构提升金融业务、运营支持、风险识别、决策执行和客户服务等能力。

6 资源池要求

6.1 大模型训练推理资源要求

大模型训练推理资源要求如下：

- a) 应保障生产环境大模型训练推理资源与非生产环境大模型训练推理资源的逻辑和物理隔离；
- b) 应保障大模型训练推理资源的基本安全，包括但不限于网络安全、主机安全、容器安全；
- c) 大模型训练推理资源应支持弹性与横向扩展能力；
- d) 应确保通信带宽满足训推需求，包括但不限于构建RDMA高速互联网络或节点内高速互联网络；
- e) 大模型训练推理资源应符合GB/T 45288.1中关于资源池的要求。

6.2 大模型训练推理资源虚拟化要求

大模型训练推理资源虚拟化要求如下：

- a) 大模型训练推理资源应通过虚拟化方式实现资源的统一管理，包括但不限于调度、回收、监控、运维；
- b) 资源虚拟化不应影响生产环境与非生产环境训练推理资源的逻辑和物理隔离；
- c) 虚拟化应支持按照人工智能加速卡数量进行实例规格划分；
- d) 虚拟化应支持人工智能加速卡的显存及算力切分；
- e) 应保障实例之间的安全隔离，实例应至少为用户级。

6.3 调度策略要求

6.3.1 任务分级机制

任务分级机制要求如下：

- a) 宜建立任务分级机制，明确任务分级的适用范围、分级维度、分级规则和管理流程；
- b) 应至少根据业务重要性、时效要求、风险影响等因素，对任务进行优先级划分；
- c) 所建立的任务分级机制应能够有效保障关键核心业务不受影响。

6.3.2 资源调度策略

资源调度策略要求如下：

- a) 应实现负载均衡等动态资源调度策略；
- b) 若建立任务分级机制，应通过优先级配置、资源预留、资源抢占等方式，优先保障关键核心业务稳定性。

7 金融优化训练数据要求

7.1 金融优化训练数据通用要求

金融优化训练数据通用要求如下：

- a) 在收集金融优化训练数据用于金融大模型优化训练前，应明确金融优化训练数据的来源、构建方式、数量级以及整体不良信息率等要求；
- b) 金融优化训练数据类型应至少包含质量优化训练数据以及对抗性数据，以提高金融大模型对特定应用的处理能力和鲁棒性，还可包含用于理解相关安全合规要求、识别金融不良信息的对齐数据；
- c) 应明确质量优化训练数据样本要求，以保障质量优化训练数据在内容、数据结构、文件格式、多样性等方面与大模型金融应用处理场景一致；
- d) 应建立统一的不可信金融优化训练数据来源库，对在数据的收集阶段、准入评估阶段、使用阶段等过程中发现的不可信数据源进行记录，避免再次采集及使用。

7.2 金融优化训练数据来源要求

7.2.1 开源数据

开源数据作为金融优化训练数据来源的要求如下：

- a) 应遵守开源授权协议或相关授权文件；
- b) 应优先选择权威性强、行业认可度高的开源数据集；
- c) 选择权威性或行业认可度不高的开源数据集时，应通过统计分析、异常检测、人工查验等方式评估并保障所选金融优化训练数据质量。

7.2.2 自采数据

自采数据作为金融优化训练数据来源的要求如下：

- a) 应对采集来源、采集方式、采集用途等信息进行记录；
- b) 应对自采数据中的核心数据、重要数据或敏感数据进行匿名化等预处理操作；
- c) 若采集来源为互联网数据，不应采集他人已明确不可采集的信息；
- d) 若采集来源为互联网数据，应优先选择权威性强、行业认可度高的数据来源；
- e) 对于数据可信度不高的数据来源应先进行人工审查，若存在较多可疑或不可信数据不应进行采集，不可信数据来源宜纳入不可信数据来源库中；
- f) 若需对采集后的数据进行标注处理，应符合GB/T 45674的要求；
- g) 应对标注数据的语料来源、标注方式、使用用途等信息进行记录。

注：声明不可采集的方式包括但不限于robots协议等。

7.2.3 生成合成数据

生成合成数据作为金融优化训练数据来源的要求如下：

- a) 若大模型金融应用处理的输入数据能够基于规则进行生成合成的，可基于生成合成方式构建金融优化训练数据；
- b) 应明确生成合成数据规则，包括字段类型、字符集、随机占比、处理逻辑、约束条件等规则要求；
- c) 不应将核心数据、重要数据、敏感数据引入到生成合成数据中。

7.2.4 商业数据

商业数据作为金融优化训练数据来源的要求如下：

- a) 采购或以合作形式获取第三方数据前，应要求数据提供方或数据处理方对金融优化训练数据或产品的来源合规性、质量等方面作出承诺并提供相关证明材料；

- b) 应对数据提供方或数据处理方所提供的证明材料进行审核，未通过审核的不应使用；
- c) 在使用商业数据前，应先签订具有法律效力的交易合同、合作协议等材料；
- d) 商业数据的使用不应超过合同、协议的使用范围。

7.3 金融优化训练数据准入控制要求

金融优化训练数据准入控制要求如下：

- a) 应建立合理的准入控制机制，对金融优化训练数据实施准入控制；
- b) 准入控制机制至少包含金融优化训练数据入库前的数据质量及安全评估，评估内容及要求如下：
 - 1) 评估质量优化训练数据与大模型金融应用在内容、格式、全面性、正负样本分布等方面的一致性；
 - 2) 通过数据统计分析、异常数据检测、人工审查等方式评估质量优化训练数据是否存在重复、矛盾、偏见、歧视性、误导性、不良信息等影响金融大模型优化训练效果的有害数据；
 - 3) 评估质量优化训练数据是否存在核心数据、重要数据、敏感数据以及泄露后会造成业务合规风险的数据；
 - 4) 若应用于智能客服等公共服务类场景，应评估是否满足相关监管机构及行业要求。

7.4 金融优化训练数据安全及管理维护要求

入库后的金融优化训练数据安全及管理要求如下：

- a) 应建立安全有效的金融优化训练数据访问控制机制并遵循最小权限原则；
- b) 应对入库后的金融优化训练数据增删改查等操作进行严格管控，避免金融优化训练数据被恶意篡改、删除或泄露；
- c) 应定期（至少每年一次）对金融优化训练数据集进行审查及维护，确保金融优化训练数据的质量和适宜性。

8 金融应用大模型要求

8.1 基础大模型选择要求

基础大模型选择要求如下：

- a) 应优先选择已备案或可信机构发布的大模型作为基础大模型；
- b) 若使用外部基础大模型用于智能客服等公共服务类应用场景，应选择已备案大模型；
- c) 所选基础大模型应具备良好的输入信息理解及指令遵循能力，具备理解输入信息并根据格式约束、语义约束等约束条件进行信息处理的能力；
- d) 所选基础大模型通过系统提示词优化、知识库配置等工程手段部署后，准确率与预期目标差距较大时，不应作为金融大模型优化的基础大模型；
- e) 若大模型金融应用需要使用广泛的金融知识、数学知识、深度推理能力、复杂任务拆解能力、基于超长上下文的理解等增强能力时，应对基础大模型所需的专项能力进行评估或参考权威的评估结果，以确保所选的基础大模型具备良好的金融应用处理能力；
- f) 若大模型金融应用输入数据不可控，没有人工录入或审核等环节，所选基础大模型要求如下：
 - 1) 应具备良好的鲁棒性，能够识别和处理用户输入中的错别字、语法错误等异常数据；
 - 2) 具有抵御常见的提示注入等攻击能力。
- g) 在基础大模型处理能力满足要求的前提下，应优先选择参数量小的大模型以降低推理成本、提高响应速度，进一步保障业务稳定性需求；

- h) 应持续或定期（至少半年一次）关注可选基础大模型，当可选基础大模型处理能力有重大提升时，应重新评估待开发大模型金融应用的可行性；
- i) 使用开源模型或商业模型作为基础大模型前，应进行合规性审查，防止在超出许可约定范围或未经授权的情况下使用。

8.2 金融大模型优化要求

金融大模型优化要求如下：

- a) 应使用入库后的金融优化训练数据进行金融大模型优化训练；
- b) 应关注金融优化训练数据对金融大模型处理效果的影响；
- c) 在训练推理资源紧缺的情况下应在空闲时段进行金融大模型优化训练，以降低对大模型金融应用的影响；
- d) 应对金融大模型所使用的金融优化训练数据版本信息进行记录，便于追溯和审计。

8.3 金融应用大模型准入控制要求

应对金融应用大模型实施准入控制，金融应用大模型是否满足金融应用准入要求应根据11.1的大模型金融应用准入控制要求进行评估。

8.4 金融应用大模型安全管理要求

金融应用大模型安全管理要求如下：

- a) 应对准入后的金融应用大模型文件进行安全存储并建立安全有效的访问控制机制，避免金融应用大模型原始文件被恶意篡改或泄露；
- b) 应通过完整性校验等机制保障金融应用大模型在部署运行阶段的完整性和一致性，防止部署运行阶段被恶意篡改或泄露；
- c) 应建立模型卡，对模型基本信息、评估结果、使用的金融优化训练数据、采用的优化策略、金融应用情况等信息进行详细记录。

8.5 金融应用大模型升级更新要求

金融应用大模型升级更新要求如下：

- a) 当对已准入的金融应用大模型进行优化训练或修改配置参数等操作时，应重新进行金融应用大模型准入控制流程；
- b) 金融应用大模型的升级更新及部署不应影响金融业务处理造成中断性等影响，若无法避免影响，应采用人工接入等方式降低影响。

9 系统组件要求

9.1 协议接口要求

协议接口要求如下：

- a) 应建立统一的大模型金融应用接口平台，用于进行接口的新增、审核、访问控制及删除，以实现接口的统一管理；
- b) 应对接口的使用情况进行监控，若发现异常调用情况应采取安全措施；
- c) 不应通过API等方式直接调用外部服务，也不应向外部机构直接提供服务接口；

- d) 若采用智能体等由模型自主选择并调用接口服务的方式，宜使用MCP等统一接口协议以加强各类大模型对接口的理解和调用能力。

9.2 知识库要求

9.2.1 知识库内容要求

知识库内容要求如下：

- a) 应支持数据库作为知识数据源；
- b) 应支持结构化文档作为数据源，包括但不限于xls、csv文件；
- c) 知识库增添数据应满足最小必要原则，只上传大模型金融应用所需的最少数据集，以保障检索效率及准确性；
- d) 应定期（至少半年一次）对知识库内容进行审查及维护，以保持知识库内容的良好适宜性和时效性；
- e) 若知识库对应的大模型金融应用不再使用，应及时删除相关知识库；
- f) 应对添加到知识库的数据进行预处理，删除无意义、非必要的冗余信息，确保知识内容的精炼，通过结构化及标签等方式，提高知识库检索准确率及效率。

9.2.2 知识库构建要求

知识库构建包括以下内容：

- a) 知识库应支持常见文档格式解析，如pdf、txt、doc、docx、xlsx、ppt、pptx、eml、jpg、png、csv、md等；
- b) 嵌入模型的选型应考虑其在金融领域知识表征的准确性，所选的嵌入模型应具备良好的语义理解能力，能够有效捕捉金融术语、语义关系及差异性；
- c) 应通过多路召回、重排序等方式，保障知识库检索效果；
- d) 对于实体和关系复杂的知识，宜通过知识图谱等技术进行构建，以保障深层次语义关联检索的准确性，提升复杂金融场景下的问答和推理能力。

9.2.3 知识库安全要求

知识库安全要求如下：

- a) 应在知识库内容添加前进行内容审核，确保新增内容不包含不良及虚假信息；
- b) 知识库应实现知识的精细化隔离及访问控制，确保机构内外部使用者无法通过知识库直接访问或通过金融应用大模型间接检索等方式，获得原有渠道无法获得的信息；
- c) 应建立知识内容溯源机制，记录每一条知识的来源、创建者、修改历史、审核记录和发布时间，以便于追溯和审计；
- d) 在建立知识库的过程中，若涉及个人信息的收集、使用、存储和共享，应符合GB/T 35273、JR/T 0171的相关要求；
- e) 若用于智能客服等金融公共服务类场景，知识库中的数据不应包含核心数据、重要数据以及敏感数据。

9.3 标识工具要求

标识工具要求如下：

- a) 若用于智能客服等金融公共服务类场景，标识工具应用效果应符合GB 45438的要求；

- b) 若大模型金融应用的生成合成内容与非生成合成内容在流程中难以人工辨别,应进行显式提醒,以便于输出内容的人工审核及验证;
- c) 显式标识及隐式标识工具的可靠性、性能应经过验证并满足业务处理场景需求。

9.4 关键词库及安全风险验证库要求

关键词库及安全风险验证库要求如下:

- a) 若用于智能客服等金融应用大模型输入数据不可控的场景,应通过关键词库及安全风险验证库,加强模型输入侧的攻击抵御能力;
- b) 关键词库应围绕GB/T 45654规定的训练数据及生成内容安全风险类别,结合各类风险的典型表现形式和金融业务场景,建立具有代表性和适用性的风险关键词、短语或规则;
- c) 应保障违法不良关键词库及安全风险验证库的数据安全,避免相关数据泄露后造成的特定攻击绕过风险;
- d) 应对关键词库进行分类管理,类型可分为敏感词、风险提示词、业务引导词、禁用词等;
- e) 应对关键词触发后的响应机制进行分级,可根据其风险程度设定不同响应级别,如“拒绝回答”“警示提醒”“转人工服务”等;
- f) 应建立关键词库及安全风险验证库更新机制,定期(至少每月一次)审核更新的必要性,在金融政策重大调整等情况下,应及时完成更新;
- g) 更新方式包括内部来源更新及外部来源更新:
 - 1) 内部来源更新主要是对于未能通过原有关键词库及安全风险验证库等技术检测识别到的非法输入及输出内容,应经过个人信息脱敏处理后纳入安全风险验证库中,若涉及关键词触发的安全风险,应纳入关键词库中;
 - 2) 外部来源更新主要是根据金融市场、监管要求、可信公开渠道等信息,进行必要的关键词库及安全风险验证库更新。
- h) 关键词的定义与匹配规则应明确、精准,避免因泛化或歧义导致的高误报率,影响正常用户交互;
- i) 应保障关键词库及安全风险验证库的全面多样性,能够满足各类大模型金融应用需求;
- j) 通用场景及特定场景的关键词库及安全风险验证库应分开管理维护;
- k) 对关键词库的所有增、删、改操作均应记录日志,日志内容应包括操作人、操作时间、变更内容及审批记录,确保所有变更可审计、可追溯。

10 管控服务平台要求

10.1 管控服务平台功能要求

管控服务平台功能要求如下:

- a) 管控服务平台可由多个服务平台组成或构建统一的管控服务平台,平台应实现对资源池、金融优化训练数据、金融应用大模型、系统组件及大模型金融应用的统一管控,至少能满足本文件7.3、8.3、11.1的准入控制要求、7.4、8.4、8.5、11.2及11.4的安全管理、维护及升级更新要求;
- b) 管控服务平台应提供开源数据、开源大模型、开源知识库、开源标识工具等大模型开发、部署、应用所涉及的开源软件管控服务,应满足行业关于开源软件引入、使用、管理等相关要求;
- c) 应对所管控对象的关键操作进行日志记录,日志内容至少包括操作主体、操作对象、操作类型、操作时间、操作结果等信息;

- d) 不应提供绕过准入控制、安全管理、维护及升级更新流程的功能。

10.2 管控服务平台安全要求

管控服务平台安全要求如下：

- a) 管控服务平台的使用及管理均应经过身份认证，不应使用默认或弱密码；
- b) 管控服务平台的关键参数配置需经审核后，由管理人员进行操作；
- c) 不应使用默认或非必要端口，防止外部攻击风险；
- d) 应支持对异常调用、越权访问、敏感信息泄露、模型输出风险等情况进行监测和告警。

11 大模型金融应用要求

11.1 大模型金融应用准入控制要求

应建立合理的大模型金融应用准入控制机制，准入前应经过金融机构内部或第三方机构评估，评估内容应至少包括输出内容准确性、输出内容可解释性、金融应用安全性以及金融应用处理性能，所涉及的金融应用大模型、知识库、系统提示词、管控服务平台、验证数据类型等大模型金融应用评估环境应与实际保持一致。

11.1.1 输出内容准确性评估

输出内容准确性评估要求如下：

- a) 验证数据集不应从金融优化训练数据集中直接抽取，可在构建金融优化训练数据集时同步构建验证数据集；
- b) 所构建的验证数据集应具备完整性、正确性、多样性，应通过人工审查等方式保障验证数据集质量；
- c) 应根据大模型金融应用类型使用合理的评价指标，通过多个指标进行综合分析以发现金融应用处理缺陷，例如分类任务指标可包括准确率、召回率、精确率、F1分值等；
- d) 应根据大模型金融应用类型建立合理的准入基准，例如分类任务中非资金类大模型金融应用输出内容准确率应大于90%，资金类大模型金融应用输出准确率应大于95%；
- e) 对于无法通过评价指标进行评价的大模型金融应用，应通过裁判大模型、人工评价、专家评审等方式确定大模型金融应用是否符合准入要求。

11.1.2 输出内容可解释性评估

当大模型金融应用的输入数据及输出结果之间不存在显而易见的因果关系或难以通过人工审核直接判断时，金融应用大模型的输出结果至少包括：

- a) 推理过程描述；
- b) 推理及决策依据文件描述；
- c) 决策合理性解释；
- d) 其他决策的不合理性解释；
- e) 决策公平性解释。

11.1.3 金融应用安全性评估

若应用于智能客服等公共服务类场景，大模型金融应用安全性评估内容包括：

- a) 输出内容不应包含核心数据、重要数据以及敏感数据；

- b) 应对可能导致公众混淆或者误认的生成合成内容添加显式标识,提供生成合成内容文件的应添加隐式标识,标识效果应符合GB 45438的要求;
- c) 应根据大模型金融应用形式完成生成式人工智能服务备案或服务登记;
- d) 应满足监管机构及行业其他要求。

11.1.4 金融应用处理性能评估

应对大模型金融应用处理性能进行评估,大模型金融应用处理性能准入基准应满足 11.5 的大模型金融应用处理性能要求。

11.2 大模型金融应用安全管理要求

11.2.1 大模型金融应用通用安全管理要求

大模型金融应用通用安全管理要求如下:

- a) 应建立有效的大模型金融应用安全管理机制,未经准入控制不应部署使用;
- b) 应对准入后的金融应用大模型、系统组件、系统提示词、配置信息等影响大模型金融应用效果的软件、模型、配置信息进行记录、安全存储并建立安全有效的访问控制机制,避免被恶意篡改或泄露;
- c) 应通过完整性校验等机制保障大模型金融应用在部署运行阶段中的完整性和一致性,防止部署运行阶段被恶意篡改或泄露;
- d) 不应使用默认或非必要端口;
- e) 应保障所能调用的工具及服务不存在恶意代码或行为。

11.2.2 智能体金融应用安全管理要求

智能体金融应用安全管理要求如下:

- a) 权限申请应遵循最小必要原则,符合大模型金融应用的安全管理和访问控制逻辑,不申请、使用与任务无关的权限,仅访问其完成当前任务所必需的文件、服务和工具;
- b) 不应授权智能体通过工具调用、接口协议等任何方式直接进行敏感操作,敏感操作包括:
 - 1) 对金融业务等原始数据或文件进行修改、删除等操作;
 - 2) 对系统、平台、组件、大模型等配置信息进行添加、修改、删除等操作;
 - 3) 执行脚本、软件等可运行程序;
 - 4) 模拟用户点击;
 - 5) 影响金融业务系统或金融业务处理的其他操作。
- c) 当智能体需要进行敏感操作时,操作内容应经过严格的人工审查,确认无误后进行;
- d) 应针对智能体授权的操作范围,建立有效的恢复机制,在错误操作后能够立即恢复;
- e) 应定期审查(每年至少一次)并建立有效的智能体安全风险策略,在出现风险或策略不允许时应能有效停止调用或人工审查接管;
- f) 不应利用系统权限自动化读取、操作应用的图形用户界面或模拟用户点击,所有交互应通过标准化接口进行。

11.3 大模型金融应用运行要求

11.3.1 输入内容审查

输入内容审查要求如下:

- a) 当大模型金融应用会对金融业务处理结果或金融消费者权益产生影响时,应对大模型金融应用输入内容质量及安全进行审查,以降低内外部攻击风险,提高输出内容可信度;
- b) 输入内容质量审查应包括输入内容的完整性及真实性、数据结构的正确性、文件格式的符合性等质量审查点;
- c) 输入内容安全审查应包括是否含有注入攻击特征、不良违法信息、篡改痕迹、恶意意图等安全审查点;
- d) 为提高审核效率,可通过注入攻击识别、不良违法信息检测、文件格式校验等自动化检测方式辅助判定,但最终的审核结果应交由人工审核及判断并保留审查人员、审查结果等信息;
- e) 若输入内容来自机构内部人员录入等已经过审查的可信渠道,质量审查及安全审查点应在人工录入等上一环节中进行;
- f) 人工审查发现质量或安全问题的,应采取人工修改、中断等处理方式并保留相关处理日志,防止恶意攻击或不符合要求的输入内容传输至金融应用大模型;
- g) 在人工审查过程中若发现关键词库或安全风险验证库等技术检测未能识别到的注入攻击、恶意意图等非法输入内容时,应进行记录并定期(至少每月一次)更新至关键词库或安全风险验证库。

11.3.2 输出内容使用

输出内容使用要求如下:

- a) 若大模型金融应用对金融业务处理有影响时,金融应用大模型输出内容仅能作为数据处理的中间参考结果用于提高人工处理效率,最终输出内容或决策应由人工决定;
- b) 大模型金融应用输出内容作为中间参考结果应重点关注内容质量及安全;
- c) 输出内容质量应关注输出内容完整性、数据结构正确性、文件格式符合性等检查点;
- d) 输出内容安全应关注是否含有不良违法信息、虚假信息、偏见及歧视性、逻辑错误等检查点;
- e) 为提高内容检查效率,可通过不良违法信息检测、输出内容完整性校验、文件格式校验等自动化检测方式辅助判定;
- f) 若检查无误后可直接采纳输出内容;
- g) 若发现输出内容存在质量或安全问题,应采取人工修改、中断等处理方式,防止输出内容对后续业务流程造成影响;
- h) 应对检查过程中发现的大模型金融应用输出质量及安全问题进行记录,为大模型金融应用优化分析及升级更新提供数据支撑。

11.3.3 输出内容监控

输出内容监控要求如下:

- a) 若大模型金融应用对金融业务处理有影响时,应对金融应用大模型输出内容整体表现进行监控;
- b) 应对大模型金融应用处理准确率、安全风险等关键指标进行监控;
- c) 应对大模型金融应用处理准确率、安全风险等监控指标设置预警线,若触发预警线,应及时采取相应措施、如大模型金融应用进行下架、分析、优化处理,并根据11.4的大模型金融应用升级更新要求进行升级更新。

11.3.4 运行稳定性保障

运行稳定性保障要求如下:

- a) 若对金融业务处理有影响时,应保障大模型金融应用运行的稳定性;
- b) 资源池应能够保障关键大模型金融应用运行的稳定性;

- c) 当出现大模型金融应用中断时应及时通过拉取等方式进行恢复;
- d) 应建立大模型金融应用备选方案, 在应用处理无法及时恢复时采取应急措施。

11.4 大模型金融应用升级更新要求

大模型金融应用升级更新要求如下:

- a) 应建立大模型金融应用升级更新管控机制, 保障在受控情况下进行升级更新;
- b) 当大模型金融应用需要进行系统组件升级更新、金融应用大模型升级更新、系统提示词优化、配置信息修改等影响大模型金融应用效果的软件、模型、配置信息更新时, 应重新进行大模型金融应用准入控制;
- c) 大模型金融应用升级更新前应做好相关备份, 以便在发生异常时能够恢复或回退;
- d) 大模型金融应用的升级更新及部署不应影响金融业务处理造成中断性等影响, 若无法避免影响应采用人工接入等方式降低影响。

11.5 大模型金融应用处理性能要求

大模型金融应用处理性能要求如下:

- a) 应依据GB/T 45225和JR/T 0221中关于性能评价的要求, 评测大模型金融应用处理性能, 评价指标应至少包括TTFT、TPOT、E2E Latency、吞吐量(QPS/TPS)等;
- b) 大模型金融应用处理性能指标应满足金融业务处理需求。

参 考 文 献

- [1] GB/Z 185.1—2026 人工智能 智能体互联 第1部分：总体架构
 - [2] GB/T 9813.3—2017 计算机通用规范 第3部分：服务器
 - [3] GB/T 45288.2—2025 人工智能 大模型 第2部分：评测指标与方法
 - [4] GB/T 45288.3—2025 人工智能 大模型 第3部分：服务能力成熟度评估
 - [5] JR/T 0197—2020 金融数据安全 数据安全分级指南
 - [6] 国家金融监督管理总局。银行保险机构数据安全管理办法[S]. 金规(2024)24号, 2024-12-27.
-